LIMSI at MediaEval 2015: Person Discovery in Broadcast TV Task

Johann Poignant, Hervé Bredin, Claude Barras
LIMSI - CNRS - Rue John Von Neumann, Orsay, France.
firstname.lastname@limsi.fr

ABSTRACT

This paper describes the algorithm tested by the LIMSI team in the MediaEval 2015 Person Discovery in Broadcast TV Task. For this task we used an audio/video diarization process constrained by names written on screen. These names are used to both identify clusters and prevent the fusion of two clusters with different co-occurring names. This method obtained 83.1% of EwMAP tuned on the out-domain development corpus.

1. INTRODUCTION

We present the approach of the LIMSI team to the Person Discovery in Broadcast TV Task at MediaEval 2015. To address this task we had to return the names of people who can be both seen as well as heard in a selection of shots in a collection of videos. The list of people is not known *a priori* and their names must be discovered in an unsupervised way from media content using text overlay or speech transcripts. For further details about the task, dataset and metrics the reader can refer to the task description [4].

We first detail the fusion system baseline provided to all participants (we are both organizer and participant). Then, we describe the constrained multi-modal clustering. Finally, we compare the results obtained.

2. MULTI-MODAL FUSION

We propose two different approaches to address the task. They only rely on metadata provided to all participants (see Table 1). Only written names are used as source of identity. In addition to speech turn segmentation and face detection and tracking, the baseline relies on the provided speaker diarization and speaking face mapping. The constrained clustering relies on the similarity matrices (for speaker and face) to process its own clustering.

2.1 Baseline

From the written names and the speaker diarization, we used the "Direct Speech Turn Tagging" method described in [5] to identify speaker: we first tagged speech turns with co-occurring written name. Then, on the remaining unnamed speech turns, we find the one-to-one mapping that maximizes the co-occurrence duration between speaker clusters and written names (see [5] for more details). Finally, we propagate the speaker identities on the co-occurring face tracks based on the speech turns/face tracks mapping.

Components	Baseline	Constrained clustering
Speech turns		
Segmentation	x	x
Similarity		x
Diarization	x	
Face		
Detection & Tracking	x	x
Similarity		x
Diarization		
Speaking face		
Mapping	x	x
Source of names		
Written names [3]	x	x
Pronounced names [2, 1]		

Table 1: Sub-component provided used by fusions

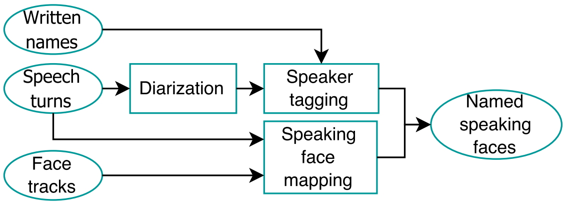

Figure 1: Baseline fusion system overview

2.2 Constrained multi-modal clustering

Figure 2 shows a global overview of our method. We first combined the mono-modal similarity matrix and the speaking face mapping into a large multi-modal matrix using weights α and β to give more or less importance to a given modality. In parallel, written names are used to identify co-occurring face tracks and speech turns.

Then, we perform an agglomerative clustering on the multimodal matrix to merge all face tracks and speech turns of a same person into a unique cluster. This process is constrained by avoiding the fusion of clusters named differently. The two parameters α and β advance or delay the merge of components of a modality relatively to others during the agglomerative clustering process, while the stopping criterion is chosen to maximize the target metrics (here the EwMAP).

MediaEval 2015 Workshop Sept. 14-15, 2015, Wurzen, Germany

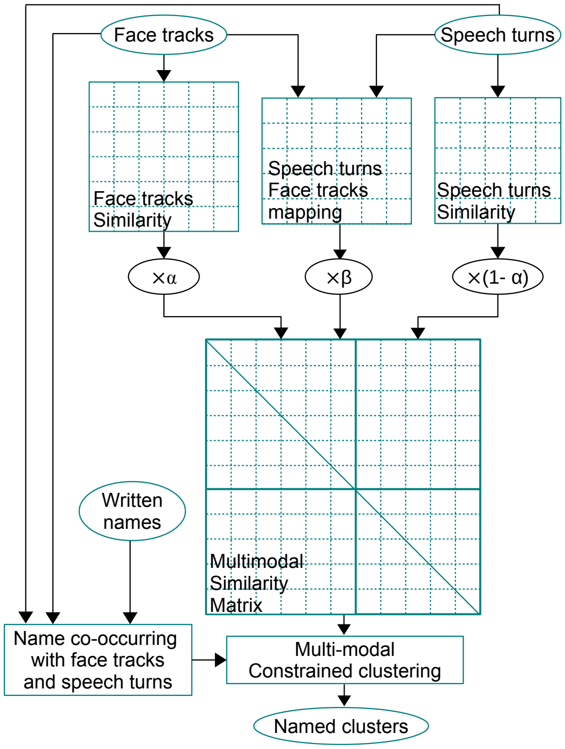
Face tracks
Speech turns
Face tracks Similarity
Speech turns Face tracks mapping
Speech turns Similarity
×α
×β
×(1- α)
Written names
Multimodal Similarity Matrix
Name co-occurring with face tracks and speech turns
Multi-modal Constrained clustering
Named clusters

Figure 2: Constrained clustering overview

A complete description of this method can be found in [6].

2.3 Speaking face selection and confidence

The last part is common for the two fusions. For each person who speaks and appears in a shot (following the shot segmentation provided to all participants), we compute a confidence score. This score is based on the temporal distance between the speaking face and its closest written name. This confidence equals to:

$$\text{confidence} = \begin{cases} 1+d & \text{if the speaking face co-occurs} \\ & \text{with the written name} \\ 1/\delta & \text{otherwise} \end{cases}$$

where d is the co-occurrence duration and δ is the duration of the gap between the face track (or speech turn) and the written name.

3. RESULTS

In Table 2, we report the EwMAP, the MAP and the Correctness (denoted by C) obtained by the baseline and the constrained clustering tuned on an out-domain corpus (for the first deadline: 01-jul-15) and on an in-domain corpus (second deadline: 08-jul-15).

The baseline does not take into account the similarity between face and does not benefit from the knowledge of written names during the diarization process. In addition to these 2 additional information, our second method optimizes the stopping criterion of the clustering based on the target metric (EwMAP) while the diarization of the baseline is tuned to maximize the classical DER.

For the first deadline (July 1st) we tuned the parameters α and β and the stopping criterion of the clustering process on the out-domain development set. For the second deadline (July 8th), we tuned these parameters with the evaluation proposed via the leaderboard (computed every six hours on a subset of the test set). We can see only a little improvement between them, showing that our method generalizes well.

Run	EwMAP(%)	MAP(%)	C(%)
Baseline	78.35	78.64	92.71
Const. clus. 01-jul-15	83.13	83.46	93.19
Const. clus. 08-jul-15	84.56	84.89	94.11
Oracle propagation mono-show	96.84	96.84	97.25
Oracle propagation cross-show	97.83	97.83	97.83

Table 2: Results

To determine the scope for further progress we used an oracle capable of recognizing a speaking face as soon as his/her written name is correctly extracted by the OCR module. In the mono-show case, the name must be written in the same video. In the cross-show case, the name can be written in any video of the corpus. Since our own approach only uses mono-show propagation, these oracle experiments show it is possible to earn up to 1% of MAP using cross-show propagation approaches.

In Table 3 we report the mean precision and recall over all queries. Compared to the baseline, the constraints on the clustering process allows to have a lower stopping criterion (therefore to have bigger clusters and hence to improve the recall), while keeping very good clusters purity (see the precision in Table 3). The high precision of our constraint clustering made the choice of the confidence score (used to rank shots in the computation of the MAP) not really important. The tuning of the three parameters on an in-domain corpus improves recall by 1.3% and decreases precision by 0.8%. In practice, α was reduced for the July 8th (in-domain tuning), therefore speech turns clustering was delayed (with respect to face tracks clustering) between July 1st (out-domain) and July 8th (in-domain tuning).

Run	Precision(%)	Recall(%)
Baseline	79.1	74.8
Const. clus. 01-jul-15	98.5	82.9
Const. clus. 08-jul-15	97.7	84.2

Table 3: Mean precision and recall

4. CONCLUSION AND FUTURE WORKS

This paper presented our approach and results at the MediaEval Person Discovery in Broadcast TV task. The process used an audio/video diarization constrained by written names on screen. This source of identities is used to both identify clusters and avoid wrong merges during the agglomerative clustering process.

For future works we will improve the distance between speech turns, try other clustering methods and cross-show propagation.

Acknowledgment. This work was supported by the French National Agency for Research under grant ANR-12-CHRI-0006-01 (CAMOMILE project).

5. REFERENCES

[1] M. Dinarelli and S. Rosset. Models Cascade for Tree-Structured Named Entity Detection. In *IJCNLP*, 2011.

[2] L. Lamel, S. Courcinous, J. Despres, J. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V.-B. Le, H. Ney, M. Nussbaum-Thom, I. Oparin, T. Schlippe, R. Schlëter, T. Schultz, T. F. da Silva, S. Stüker, M. Sundermeyer, B. Vieru, N. Vu, A. Waibel, and C. Woehrling. Speech Recognition for Machine Translation in Quaero. In *IWSLT*, 2011.

[3] J. Poignant, L. Besacier, G. Quénot, and F. Thollard. From text detection in videos to person identification. In *ICME*, 2012.

[4] J. Poignant, H. Bredin, and C. Barras. Multimodal Person Discovery in Broadcast TV at MediaEval 2015. In *MEDIAEVAL*, 2015.

[5] J. Poignant, H. Bredin, V. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised speaker identification using overlaid texts in TV broadcast. In *INTERSPEECH*, 2012.

[6] J. Poignant, G. Fortier, L. Besacier, and G. Quénot. Naming multi-modal clusters to identify persons in TV broadcast. *MTAP*, 2015.